Measurement for Social Network Data Currency and Trustworthiness

Fan Xiaojiang
School of Computer Science, Beijing Information Science & Technology University
Beijing, China
e-mail: 13121919396@163.com

Zheng Liwei
School of Computer Science, Beijing Information Science & Technology University
Beijing, China
e-mail: zlw@bistu.edu.cn

Liu Jianbin
School of Computer Science, Beijing Information Science & Technology University
Beijing, China
e-mail: ljb@bistu.edu.cn

***Abstract*—Along with the explosive growth of the information in Social Network Service, the research of the quality of data has become a new hot point in related research field. High quality social data can more effectively support data mining, knowledge discovery, and can provide reliable and efficient data for users. Based on the measure problems of data quality, this paper discussed the measurement of two important dimensions of data quality: currency and trustworthiness. Computing models for currency measurement of data with or without time stamp are given. And based on the currency values, a trustworthiness measurement method is also given.**

Keywords-social network service; data quality; currency; trustworthiness

I. INTRODUCTION

With Internet from birth to now have been nearly 30 years, computer technology is constantly updated. With the advent of the Internet era, the Internet began to become the main source of people's lives. Social Network Service(SNS) is in such a situation, as a kind of network information interaction platform (see, *e.g.*, [1, 2]). Broadly speaking, the concept of SNS is an information network; the nodes in it can contain participant or entity, while the edges show the relationship between them (see, *e.g.*, [3, 4]).

From the earliest chat tool QQ, to the now popular Microblog and developed online shopping platforms. Undoubtedly, these brought convenient to people's life and fast, greatly improved the quality of life of people. But many problems caused by SNS cannot be ignored, the online shopping is not true evaluation, false rumors of the Microblog information, as well as WeChat circle of friends a large number of spam advertising and purchasing, causing mistrust among users. There exists uncertainty in the data about people living of real data obtained from the SNS. So that the researches of the data in SNS become necessary.

The quality of the data impacts on every aspect of people's lives. The currency and the trustworthiness of data are two of the main dimensions of data quality measurement; they are important factors affecting the quality of data. The data with a low currency may delay people's journey, influence the query even make it all no-meaning data and cause financial loss. This makes determining the currency of data becomes necessary. It's a severe test for people in the information era to effectively and truly assess of the trustworthiness of the Internet information when we face the distrust in the SNS.

Of course, there are a lot of dimensions of quality for measuring SNS data, such as accuracy, completeness, consistency, volatility, and so on (see, *e.g.*, [5, 6]). The currency of the data determines whether the data is the latest update order or judgment of the current real-time relational data sequence for a particular task or needs (see, *e.g.*, [7, 8]). Data trustworthiness measures whether a particular data source provides credible and real data (see, *e.g.*, [9, 10]).

In general, the currency of the data can be carried out directly by the timestamps in the result of good or bad. However, many databases do not have accurate timestamps. Database may lack of timely maintenance, or we created a database without taking into account the timestamps issues and other reasons. It may lead to the timestamp data becomes unavailable or inaccurate. Therefore, this paper proposes that social network data currency can be divided into two categories, one has available timestamp to calculate data currency, another doesn't. In the case that the data have timestamps we can directly calculate the currency of the data by using the timestamps; In other cases, that the data doesn't have timestamps we can using the redundant records and currency constraints to help achieve the measurement of data currency.

By drawing current sociology and psychology, this paper puts the trustworthiness of the data generated into the credibility of familiarity and the credibility of similarity. While depending on the different importance, the similarity is divided into internal and external similarity. Based on the previous work, this paper introduces the low-end compensation and high-end compensation, and make a classification of familiarity. Then this paper gives a more accurate analysis of the trustworthiness combining the data currency.

978-1-5090-4499-3/17/$31.00 ©2017 IEEE

II. GENERAL MODEL

When it comes to the research of the topology of SNS, we often make it concrete like a point-like network topology, where the nodes represent a person or a group which proves that the body of the social network is person, and where the edges represent the information interaction between nodes which proves the main source of data comes from the interaction between people.

So the data of SNS is a collection of actor's self-organization relationship, where each user is a data source and the amount of data exponentially explosive growth. At the same time the data contains a multi-level social entity relationship, which makes data being various. And the data is in a time-continuous state because the interaction between people is in every moment.

To formalize the data, this paper considers a model $\boldsymbol{T}$= (S, R, D) as a collection of relations where S is data Sender, R is data Receiver and D is data in the interaction.

For the data Sender, this paper considers a model S= (S_{ID}, S_{time}, S_{name}, S_{addr}) as a collection of relations where S_{ID} is the Sender's ID, S_{time} is the send time, S_{name} is the name and S_{addr} is the address of the Sender. And for the data Sender, there is a model R= (R_{ID}, R_{time}, R_{name}, R_{addr}) as a collection of relations like the model of S.

For the data in the interaction this paper considers a model like D= (D_{ID}, D_{ST}, D_{RT}, D_{type}, D_{size}, …), where D_{ID} is the ID of the data, D_{ST} is the send time, D_{RT} is the receive time, D_{type} is the type and D_{size} is the size of the data. And according to the need in actual situation, we can also define other parameters, such as D_{SID} and D_{RID} means the sender and receiver ids ID, etc.

III. CURRENCY MEASURE

A. Background

The measurement of the currency of the data in SNS usually could be divided into two cases (see [8]). One is data with timestamps, based on which we can determine data current values and currency order. The other is data without timestamps on which we can determine data current values and part of currency order according to the redundant records and the currency constraints.

For the data in SNS, this paper considers a model M= (tID, t[EID], $t_i[A_1]$, $t_i[A_2]$ …, $t_i[A_n]$), a collection of relations such that (a) tID shows the ID of this record; (b) t[EID] is the ID of this entity, which can be a sender or a receiver, even a data set; (c) if t_i[EID]= t_j[EID], then the record t_i and t_j must description the same entity; (d) A_1, A_2, ⋯, A_n are attributes of the entity in the record, and $t_i[A_n]$ is the specific value of the attribute A_n.

On the base of the model M, this paper defines the *currency order*: Considering a set of data D, whose data model is M. For the attribute A, we have two records t_i and t_j. And the *currency order*, $t_i <_A t_j$, would be true when the next two conditions are true: (a) t_i[EID] = t_j[EID]; (b) t_i[A] precedes t_j[A] appears in the set D.

In different data sets, for various entities, the precedence order of the attribute values tends to depend on the actual situation. For example, the value of the attribute A_1 is proportionate to the value of the attribute A_2, and then it means that if the data is sorted based on A_1, the order of the data of A_2 must be the same. The constraint for the value of the data in a data set is called the *currency constraint*.

Currency constraint on social network dataset is particularly prominent, for example, the time when it be received must be later than sent, the time when it lost effectiveness must be later than the time when it be punished.

B. Currency Measure with Timestamp

Timestamps often present among the many attributes in social data. The most common timestamps in the data models such as the D_{ST} and D_{RT}. When there are available data timestamps, we can directly calculate the currency of the data by the timestamps. We usually can directly know the valid state of data by the timestamps and the corresponding currency constraints.

In general, the given data always exists an effective length. So some data have a short relative length, such as personal information data in social networks. And some data have a long relative length, such as people's name whose effective length of time can be achieved for decades.

In conclusion, this paper considers a formula to calculate the currency of data with timestamps as follows:

$$cur = \max\left\{0, 1 - \frac{D_{RT} - D_{ST}}{D_{ET} - D_{RT}}\right\}$$

In this function, D_{ET} is the time when it ended. When the cur is zero, it means the data lost effectiveness. When the cur is closer to 1, it proves a high currency degree. If *cur* is closer to 0, then a low currency degree appears.

For a data table, there may be many data. We can calculate the currency for every data at first step, then average or weighted average.

The effective length can be obtained directly from the data source which can be divided into two categories: (a) We can find it directly from the data source. And if the time is later than the effective length, the data would be invalid and deleted. For example, the campus network card for students, the effective time can be directly set up by the school network center for one month. (b) We can't find it directly but we can find the relevant constraints about time in the information given in the data, such as the publish time and the failure time.

C. Currency Measure without Timestamp

In the case of missing or having inaccurate timestamps of data, there is no obvious constraints of time. We will be unable to determine whether the data state at this time has expired. In this case it usually needs the help of set of redundant records to be judged on the currency of the data.

This paper defines T_{A_i} as the latest set of records in attribute A_i. Then any record in TA_i will be a possible latest value for the attribute A_i of entity e. Details are defined as follows:

Considering a set of data D, whose data model is M, and an entity e in D. T_{A_i} can be defined as the latest set of record in attribute A_i as long as it can meet two conditions: (a) t[*EID*] is the true ID of the entity e; (b) There is no record s has the currency order $t <_A s$.

Because of the limited role of currency constraints for each table, in the vast majority of cases there will be lots of different data in T_{A_i} , then there are more than one possible latest data values. If there are altogether cnt(T_{Ai}) kinds of different data values for attribute A_i in the record of T_{A_i} , then the degree of latest data values for attribute A_i is $1/cnt(T_{A_i})$.

We can use the function following to calculate the currency without timestamps for entity e in SNS:

$$cur_e = \frac{1}{|Attr(e)|} \times \sum_{A_i \in Attr(e)} \frac{1}{cnt(T_{A_i})}$$

In this function, $|Attr(e)|$ is the amount of the attributes in $Attr(e)$.

When the importance of each attribute data in the table is not the same, we can weight every attribute as its importance in the data table. And ownership values sum equal to one. Consider this, we can change the function above like this:

$$cur_e = \sum_{A_i \in Attr(e)} \frac{1}{cnt(T_{A_i})} \times \omega_{A_i}$$

IV. TRUSTWORTHINESS MEASURE

Trustworthiness is the degree of dependence on people or things. For example, when some people make decisions, in many cases will refer to other people's views, the more and their relationship is good, the impact of their views on their own decisions will account for a large proportion.

The same is true for SNS, and the degree of Trustworthiness among people is divided into two parts: one is the trustworthiness of the familiarity, and the other is the trustworthiness of the similarity (see [11]).

For the trustworthiness of the familiarity, people are more familiar and more confidence with familiar people in real life. And those people tend to be higher trust by whom has same interests or similar experiences with their own.

For the trustworthiness of the similarity, this paper proposes to divide the similarity to two parts, namely, external similarity and internal similarities. External similarities include person's sex, age, address, etc. And internal similarities include hobbies, personality, values, etc. The external similarity can play a greater role in the same and all of the same conditions. For example, as a new class of students, people always more easily talk about with their fellow students who is more likely to be trusted in this situation.

When the external similarity is low but internal similarity is high, we can define a special trustworthiness compensation. It can be understood like this: considering a northerner and a southerner who do not know each other with big age gap, remote living area and all other external similarities are very low, but interest and preference are the same totally. It's highly credible for the southerner about something north tourism or something custom for southerners which is said by the northerner. But if you do not introduce the low-end trustworthiness compensation, we would think that these two people who do not know each other and the trustworthiness is very low. In a case that the external similarity is large but internal similarity is small, the internal similarity plays a greater role. This paper introduces a high-end trustworthiness compensation for this case relative to the low-end compensation. That is, in the case of a large external similarity, the internal similarity is also very large. Therefore, the calculation process of the social network reliability can be reflected by the familiarity and similarity calculation, and it can also reflect the trust relationship between users in real life.

$$\mathrm{Ntr}(A,N) = (\mathrm{Ftr}(A,N) + \mathrm{Str}(A,N)) \times \frac{cur_{e_A} \times cur_{e_N}}{(cur_{e_A} - cur_e) \times (cur_{e_N} - cur_e)} \quad (cur_e > \alpha)$$

In this function, (a) Ftr (A, N) is the trustworthiness of the familiarity of the user A to target user N; (b) Str(A,N) is the trustworthiness of the similarity of the user A to target user N; (c) cur_{e_A} is the currency of the data set A and cur_{e_B} is the currency of the data set B; (d) α is a parameter which can have user-set according to the actual situation, and it can be bigger with a higher requirement about the currency of the data.

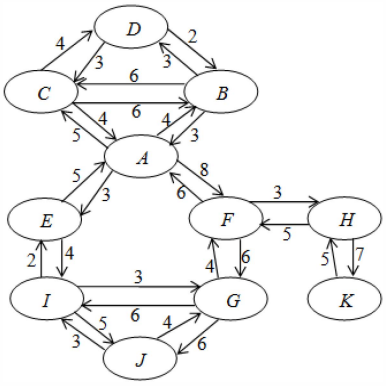

Figure 1. SNS network messages.

The general method for measuring the trustworthiness of the familiarity based on graph theory (see [4]). The method is used with some given points and lines, which can be used to describe the relationship between the certain things. Usually point to represent things, and lines to show the relationship between things: (a) in the weighted directed graph $G(N,E)$, N is the user which is the set of the node in SNS; (b) E is the exchange relationship between users, as directed edges between network nodes; (c) The connection

between the users shows that there is an exchange relationship between the two users, and the direction of the line is the recipient of the message.

As shown in Figure 1, we can see from the picture, the higher the exchange between the user, the more familiar with each other, and at the same time the exchange between the two sides is mutual. If there is a serious exchange of information between the two sides, that is, the number of A to B is 10, while B did not give A information back, which is likely to be harassed information. In this case, there is interaction between B and A, but the research of the trust between the two users is different, that is, the credibility of A to C and the credibility of C to A.

We can calculate the trustworthiness of the familiarity according to the following method: the trustworthiness of the familiarity of isolated nodes is considered as zero. In Figure 1, E and F are A's friends, G and I are J's friends. Then the path from A to J of a total of four, which are:

$$\{(A, E, I, J), (A, E, I, G, J), (A, F, G, J), (A, F, G, I, J)\}$$

For normal circumstances, the fewer points passed, the higher the reliability. According to formula (10) can be calculated confidence a function for a user of each node is:

$$\mathrm{Ftr}(A, N) = \sum_{i=1}^{n}\left(\prod_{j=1}^{m}\frac{N(S_{j-1}^{i}, S_{j}^{i})}{L_j}\right)$$

In this function, (a) A is the root node, and N is the other target nodes; (b) the parameter i is the i path in of the n path from the node A to N, and the parameter j is the layer of the path, and the parameter m is which the layer the destination node in; (c) L_j is the sum of the mix message numbers among every node in the layer j to the relative node in the layer j-1, for example, L_1=16, L_2=14, L_3=7; (d) $N(S_{j-1}^{i}, S_{j}^{i})$ is the number of the message between the node S_{j-1}^{i} and S_{j}^{i}.

Based on the research about the trustworthiness of the familiarity, we can calculate the trustworthiness of the similarity. We proposed that the similarity of users can be divided into internal and external similarity, and it can be calculated as the following function:

$$\mathrm{Str}(A, N) = \alpha S_0(A, N) + (1-\alpha) S_i(A, N)$$

In this function, (a) Str(A, N) is the trustworthiness of the familiarity between the user A and N; (b) S_0(A, N) is the internal similarity between the user A and N; (c) S_i(A, N) is the external similarity between the user A and N; (d) α is the parameter for adjustment.

When it comes to consider the external similarity, we always take account of the age and the address. They are equally important, and then assigned to the same weight. This paper proposes the function followed to calculate the external similarity:

$$S_0(A, N) = S_\alpha(A, N) + S_h(A, N)$$

In this function, S_a(A, N) is the similarity of the age, and S_h(A, N) is the similarity of the home location.

The similarity of the age will be higher when the ages are more similar. We can use another function to calculate it like:

$$S_0(A, N) = \left|Age(A) - Age(N)\right|$$

For the similarity of the home location, which is the address where the user send message in the data collected by web crawler software. In general, the more familiar with the same conditions, the closer to the home location of the user with high similarity. Here we position by using the latitude and longitude, and get the similarity of the family of home location. The closer the longitude and the latitude are, the higher the reliability is.

As the judgment for user preferences and values, This paper proposes a method of studying the internal similarity drawing on the experience of the similarity of users. This method selected the similarity measure formula, and adjusted correlation coefficient as an internal measure of similarity.

Consider a project collection $I_{A,N}$ which is given marks by the user A and N. Then the formula for the internal similarity $S_i(A, N)$ between the user A and N is:

$$S_i(A, N) = 1 + \frac{\sum_{c\in I_{A,N}}(R_{A,C} - \overline{R_A})(R_{N,C} - \overline{R_N})}{\sqrt{\sum_{c\in I_{A,N}}(R_{A,C} - \overline{R_A})^2}\sqrt{\sum_{c\in I_{A,N}}(R_{N,C} - \overline{R_N})^2}}$$

In the formula, (a) $R_{A,C}$ is the mark given by the user A for the project C; (b) $\overline{R_A}$ and $\overline{R_N}$ are the marks which are respectively given by the user A and user N. And S_i(A, N) is within the range of [0,2]. The bigger the S_i(A, N), the higher the similarity.

V. Analysis and Discussion

On the basis of the discussion and definition of the calculation of the currency, it is important to have some more complete and accurate data currency constraints, as well as the main parameters be selected in the appropriate data sets to determine the current data. The currency constraints which is defined in the decision process is to be as effective as possible, so that the decision time can be shortened to a constant level, or directly to a time constraint for a record to be effective. The most important is the definition of the currency constraint to maximize coverage records, because the higher the degree of coverage, then the impact of time constraints on the more records, which can help to better recover the data of the time series.

Secondly, the number of parameters selected in the data collection has a certain effect on the currency, but only related to the impact of the decision time. But even if it has the effect, the time is still in constant time level, so it is not done in depth.

The most difficult part of the paper is to calculate the similarity problem caused by the familiarity. Since the communication between A and the user B is directed, so we take the idea of directed graph into consideration. The similarity of familiarity is determined by the number of messages that are communicating with each other by the user.

Special emphasis is, if the number of mutual communication between *A* and *B* is serious non reciprocity, cannot be expressed as harassing messages. In the research of trustworthiness problem, it is considered that the communication of *A* to *B* is reliable and useful information, for example, user *A* order the daily special information from user *B*.

VI. CONCLUSIONS

By studying the collation and research of the existing literature, it mainly describes the definition and the method of data currency and data trustworthiness, which makes it possible to determine the currency and the trustworthiness of social network data.

In this research, the social network data timeliness is divided into two categories, one is the currency for the data with timestamps, and the other is currency for the data without timestamps. For the data with timestamps, we can directly use the dimension of time to measure the currency of the data; In the case of no timestamps, although there is some difficulty, but we found that we are still able to use the redundant records and currency constraints to measure the currency of the data. At the same time, this paper proposes to divide the research on the currency of the data without timestamps to the research for latest value and the research for currency order.

In the research, the trustworthiness of data in SNS is divided into two parts, which includes two parts, which are the trustworthiness of the familiarity and the trustworthiness of the similarity. We propose to divide the familiarity, and introduce the direction of interaction information among users, and try to compensate for the high level and low level compensation mechanism. We also put forward the measure of the reliability of the research on the basis of the timeliness of data, so that it can make the research of data credibility more realistic.

In the future, we will focus on the following issues: (a) How to measure the currency of the data after the measurement for the accuracy of the data in a data set with dynamic change; (b) How to maximize repair data through other ways when the data currency is not up to standard; (c) How to build trustworthiness model in the user's point of view, and add more factors in the standard of evaluation; (d) Combining the currency and trustworthiness of the data, how to define more methods to deal with different problems about SNS.

REFERENCES

[1] Charu C. Aggarwal. Social Network Data Analytics[C]. Springer Science+Business Media, LLC 2011.

[2] Faloutsos M, Karagiannis T, Moon S. Online social networks[J]. Network IEEE, 2010, 24(5):4-5.

[3] Han Yi, Xu Jin, Fang BinXing. Structural Supportiveness Theory on Social Networks[J]. Chinese Journal of Computers, 2014, (4):905-914.

[4] Yan Xing, Chang YaPing. A Review on the Research of Social Network Service [J]. Journal of Intelligence,2010, 29(11):44-47.

[5] Batini C, Scannapieco M. Data Quality: Concepts, Methodologies and Techniques[J]. Data-Centric systems and Applications, 2006.

[6] Han JingYu, Xu LiZhen, Dong YiSheng. An Overview of Data Quality Research[J]. Computer Science, 2008, 35(2):1-5.

[7] Fan W, Geerts F, Wijsen J. Determining the currency of data//Proceedings of the ACM Symposium on Principles of Database Systems(PODS)[J]. Athens, Greece, 2011:71-82.

[8] Li MoHan, Li JianZhong, GaoHong. Evaluation of Data Currency[J]. Chinese Journal of Computers, 2012, 35(11):2348-2360.

[9] Tseng S, Fogg B J. Credibility and computing technology[J]. Communications of the ACM, 1999, 42(5): 39-44.

[10] Gambhir M, Doja M N. Action-Based Trust Computation Algorithm for Online Social Network[C]//Advanced Computing & Communication Technologies (ACCT), 2014 Fourth International Conference on. IEEE, 2014: 451-458.

[11] Qiao XiuQuan, YangChun, Li XiaoFeng. A Trust Calculating Algorithm Based on Social Networking Service Users' Context[J]. Chinese Journal of Computers, 2011, 34(12):2403-2413.